Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network

Bernd Accou*†, Mohammad Jalilpour Monesi*†, Jair Montoya*, Hugo Van hamme† and Tom Francart*
*ExpORL, Department of Neurosciences
KU Leuven, Leuven, Belgium
†PSI, Department of Electrical Engineering
KU Leuven, Leuven, Belgium

***Abstract*—Current tests to measure whether a person can understand speech require behavioral responses from the person, which is in practice not always possible (e.g. young children). Therefore there is a need for objective measures of speech intelligibility. Recently, it has been shown that speech intelligibility can be measured by letting a person listen to natural speech, recording the electroencephalogram (EEG) and decoding the speech envelope from the EEG signal. Linear decoders are used, which is sub-optimal, as the human brain is a complex non-linear system and cannot easily be modeled by a linear decoder. We therefore propose an approach based on deep learning which can model complex non-linear relationships. Our approach is based on dilated convolutions as used in WaveNet to maximize the receptive field with regard to the number of tunable parameters. Comparison with a model based on a state of the art linear decoder and a convolutional baseline model shows that our proposed model significantly improves on both models (from 62.3% to 90.6% (p <0.001) and from 78.8% to 90.6% (p <0.001) respectively). Best results are achieved with a receptive field size between 250-500ms, which is longer than the optimal integration window for a linear decoder.**

***Index Terms*—match/mismatch, EEG decoding, speech, auditory system, envelope**

I. Introduction

A popular technique to study how the human brain processes speech is to present natural running speech to a subject and record the corresponding electroencephalogram (EEG) to capture the signal evoked by the stimulus. Then, using linear regression, either the features of the speech signal are decoded from the EEG signal (backward model), or the EEG signal is predicted from the speech signal (forward model). Finally, the correlation between the true and the predicted signal is computed, leading to a measure of neural tracking of speech [1]–[5]. This method has applications in domains such as audiology, as part of an objective measure of speech intelligibility [5]–[7]. While the results of this approach are promising, unfortunately, the correlations between actual and predicted signal with either technique are small (in the order of 0.1), limiting applicability. Additionally, when the same measurement is made multiple times, there is a large variability [4]. This is due to the use of simple linear models, which do not seem appropriate given the complexity and the dynamic nature of the brain. For instance, it is well known that depending on the level of attention and state of arousal of the subject, response latencies can change dramatically [3]. As this cannot be modeled using a pure linear approach, non-linear deep learning methods might be more suitable to tackle this task.

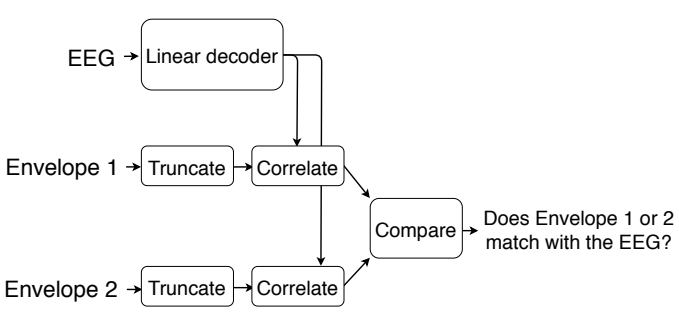
When comparing simple artificial neural networks (ANNs) to the linear decoder, somewhat higher correlations between predicted and actual signal can be obtained for EEG and for intracranial electrodes [8], [9]. For auditory attention decoding (i.e., in a multi-speaker scenario deducing which speaker a subject is attending to from the EEG signal), higher performance is reported using relatively simple deep neural networks [10]. However, the correlations between predicted and actual EEG remain low, whereas the variability across segments remains high. Therefore, long segments (in the order of several minutes) of speech and corresponding EEG are required to obtain a reliable response.

To avoid having to solve the regression problem, which is notoriously difficult with deep learning, and inspired by these recent advances in auditory attention detection, we therefore redefined the problem as a match/mismatch classification problem [11], [12]. In this paradigm, each model has 3 inputs: EEG, the corresponding stimulus envelope and an imposter envelope. This imposter envelope is a mismatched speech envelope segment. The task of the models in this paradigm is to identify the stimulus envelope corresponding to the EEG.

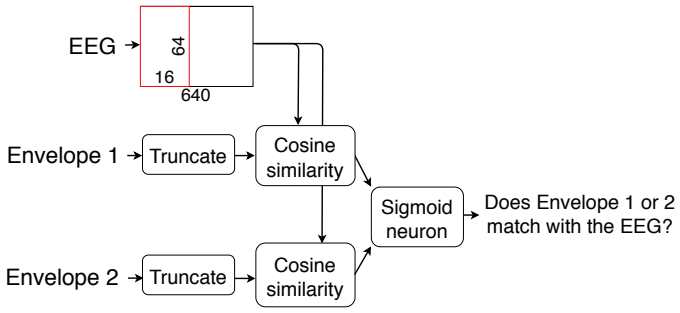
Very recently, convolutional networks have been applied for auditory attention decoding [13], [14]. Instead of a two-step approach (reconstructing the attended stimulus and comparing the similarity with the actual stimuli), these convolution-based models can classify the attended speaker directly from the EEG and the envelopes of the speech signals, which allows for end-to-end training and better performance.

Compared to fully connected ANN's, convolutional net-

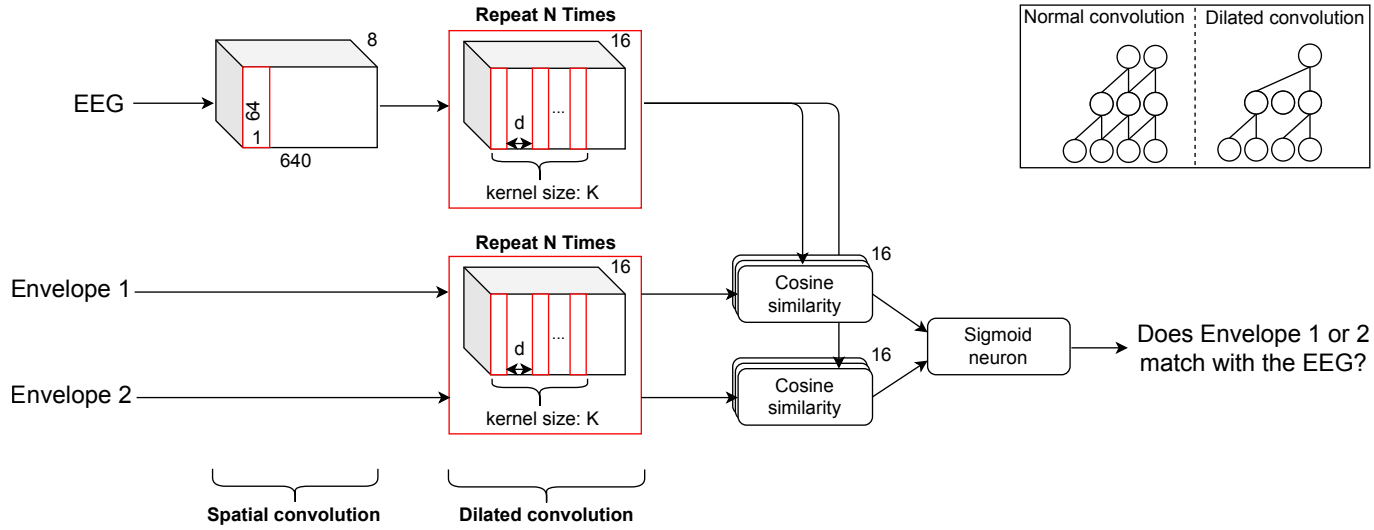
The work is funded by KU Leuven Special Research Fund C24/18/099 (C2 project to Tom Francart and Hugo Van hamme).Research funded by a PhD grant (1S89620N) of the Research Foundation Flanders (FWO).This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 637424, ERC starting Grant to Tom Francart).



(a) Decoder baseline



(b) Convolutional baseline network



(c) Dilated convolutional network

Fig. 1. The structure of the proposed networks

works are more efficient due to weight sharing, but still have limited receptive field. To maximize the receptive field of convolutional layers, while keeping the amount of parameters relatively low, dilated convolutions can be used. In dilated convolutions, some samples can be skipped by adding spacing between consecutive weights in the kernel (see Figure 1 (c)). By increasing this spacing exponentially with depth, a maximal receptive field can be achieved, while keeping the number of parameters low compared to strided convolutions. In WaveNet, dilated convolutions were used to maximize the receptive field in order to computationally cope with raw audio files at a sampling rate of 16kHz [15]. In this study, we expand on the use of dilated neural networks on time-series data by applying them to segments of EEG data and segments of speech envelopes.

II. Methods

As references, we constructed a baseline network, based on a state-of-the-art linear decoder [1], [5] as shown in Figure 1 (a) and a convolutional baseline, as shown in Figure 1 (b). To compensate for the brain delay following an auditory stimulus, both the decoder and convolutional baseline model

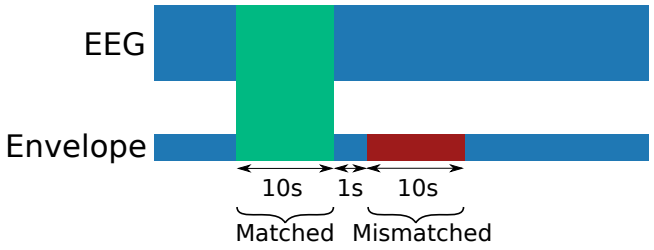


Fig. 2. To ensure similarity to the matched speech envelope segment, the imposter speech envelope segment is extracted 1 second in the future from the time aligned speech envelope segment.

apply an integration window of 250 ms to the EEG. In the decoder baseline, this integration window is constructed by concatenating the next 250 ms of EEG response to the current EEG sample in the channel dimension. The linear decoder reconstructs the envelope of the speech stimulus from the EEG by making a linear combination of all samples in the integration window. The reconstructed envelope is correlated with both input envelopes. The matched stimulus envelope is chosen based on the highest correlation value.

In the convolutional baseline, the integration window is

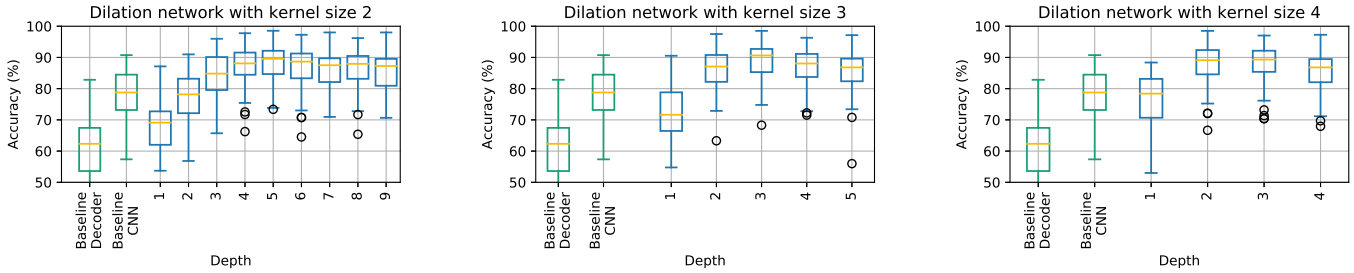


Fig. 3. Results for different kernel sizes. Each point in the boxplot is the score on the test set grouped by subject

implemented as a convolution which slides over the EEG segment and linearly combines all channels over the next 250 ms into a reconstructed speech envelope sample. The cosine similarity between the reconstructed envelope and both input envelopes is computed and fed into a single sigmoid neuron, which will classify whether input envelope 1 or input envelope 2 matches with the EEG.

Construction of the integration window for both the decoder and convolutional baseline network at the end of a segment would require samples from outside of the segment. To prevent this, the last 250 ms of each envelope segment are discarded.

Our proposed dilated convolutional network is shown in Figure 1 (c). In the first step, a convolutional layer with 8 filters is used to spatially and linearly combine all EEG channels. Then, N dilated convolutions with kernel size K are applied to the spatially filtered EEG and the two stimulus envelopes. To minimize the amount of parameters per receptive field size, the dilation factor d for layer L_n is chosen to be K^{n-1}, as in [15]. A rectified linear unit (ReLU) non-linearity is applied after each dilated convolution. Cosine similarity is used to compare each EEG representation to each stimulus representation after dilated convolutions. A sigmoid layer classifies match/mismatch based on the cosine similarity scores. The number of filters for the spatial convolutional layer and dilated convolutional layers were chosen based on the optimal performance of the model in a hyperparameter sweep (2^n filters for $n = 0...8$).

Each of the proposed models has 3 inputs, one for EEG data segments and two for speech envelope segments. Data was presented to the models in segments of 10 seconds with an overlap of 90%. Note that for linear models typically segments of 30s or longer are used. The imposter speech envelope segment is chosen 1 second after the end of the current EEG segment, as displayed in Figure 2. To ensure that our dataset is balanced, the imposter segment is alternately presented to each of the speech envelope inputs. This means that each segment of EEG is presented twice to the network.

As the only trainable weights in our baseline model are the weights of the linear decoder, the linear decoder was trained independently in a linear regression setting. In this linear regression setting, mean squared error was used as a loss function and Pearson correlation as the evaluation metric. After training, the linear decoder was evaluated as displayed in Figure 1 (a) with accuracy as the performance metric.

The convolutional baseline and dilated models were trained in an end-to-end fashion for 50 epochs with stochastic gradient descent using the Adam optimizer with a learning rate of 10^{-3}. The models were saved after each epoch. After 50 epochs, the best model was chosen from the saved models based on validation loss.

A grid search for different values of kernel size (2, 3 and 4) and depth (up to the maximal depth for each kernel size) is performed on the dilated network to search the optimal values. Depth and receptive field size are limited by the input segment length and kernel size. The maximum depth is $\lfloor \log_K(\text{segment length}) \rfloor$, e.g. for a network with kernel size 3 and a segment length of 640 samples, the maximum depth is 5, which corresponds to a receptive field size of $3^5 = 243$ samples. All models were created in tensorflow (1.14.0) [16] with the keras API [17]. The code used to construct the models is available at https://github.com/exporl/eeg-matching-eusipco2020.

III. Evaluation

To evaluate our models, we recorded EEG data from 48 normal hearing subjects while they listened to audiobooks narrated in Flemish (Dutch). Subjects were screened for normal hearing with a pure-tone audiogram and Flemish MATRIX-test. Stimulus audio was presented binaurally at 62 dBA. Different stimuli were presented per subject, chosen from a set of 10 unique stimuli of roughly the same length (14 minutes and 29 seconds $\pm$ 1 minute and 7 seconds). 23, 20, 4 and 1 subjects listened to 8, 7, 6 and 2 stimuli respectively. The order of presentation was randomized for each subject.

Each recording was split into a training, validation and test set containing 80%, 10% and 10% of the recording respectively (see Figure 4). The validation and test set were extracted from the middle of the recording to avoid possible edge effects. All presented models are trained with data across subjects to obtain general subject-independent models.

The EEG signal and speech stimulus were preprocessed in MATLAB. The envelope of the stimulus was estimated using a gamma-tone filterbank with 28 sub-bands. The envelope for each sub-band was estimated by taking the absolute value of each sample and raising it to the power of 0.6. All sub-bands were averaged to obtain 1 speech envelope [18]. EEG

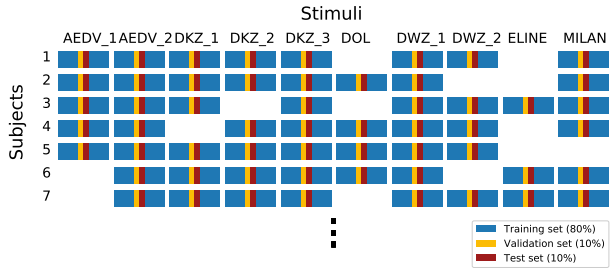


Fig. 4. Visualisation of data of 7 first subjects within the dataset. Each rectangle represents a recording.

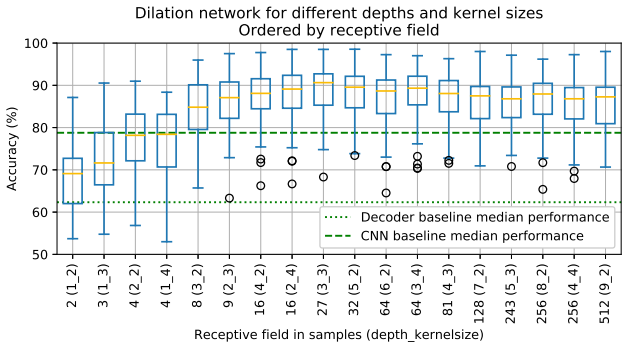


Fig. 5. Results for dilation models with different depths and kernel sizes, ordered by the receptive field size.

and speech envelopes were downsampled to 1024Hz. A Multi channel Wiener filter was applied to the EEG for artifact rejection [19]. The channels of the EEG were re-referenced to a common average. Both signals were filtered between 0.5 and 32 Hz with a Chebyshev2 filter (order of 2000, 80 dB stopband attenuation and 1 dB passband ripple) and downsampled to 64 Hz.

The stimuli were presented using a laptop running Windows using the APEX 4 software platform [20] developed at ExpORL, an RME Multiface II sound card and Etymotic ER-3A insert phones which were electromagnetically shielded. The experiments took place in an electromagnetically shielded and soundproofed cabin. The EEG was measured with an Active-Two system from BioSemi, with 64 electrodes at 8kHz sampling rate.

IV. Results

Classification accuracy for different kernel sizes K (2, 3 and 4) and depth N are shown in Figure 3. The best configurations of the dilated model outperform the convolutional baseline with 10.8%, 11.9% and 10.6% for kernel size 2, 3 and 4 respectively. All dilated convolutional models presented here outperform the decoder baseline significantly ($p < 0.001$) (Wilcoxon signed rank test with Holm-Bonferroni correction for multiple comparisons, two tailed, grouping based on subject). The dilated convolutional models also significantly outperform the convolutional baseline for depths bigger than 2 for kernel size 2, and for depths bigger than 1 for kernel size 3 and 4 (all p <0.001). For all 3 kernel sizes, we see an initial increase in match/mismatch accuracy with increasing depth to an optimum. After this optimal depth, an increase in depth has decreased performance.

The combined results in function of receptive field are displayed in Figure 5. Performance increases as receptive field size increases up to 27 samples (= 420 ms). Increasing the receptive field beyond 27 samples lowers performance with approximately 3%.

V. Discussion

The decoder baseline in our setup has a median performance of 62.3%, which is close to chance level. This low score is partly due to the wider frequency range (0.5 - 32Hz) than the one typically used (0.5 - 4Hz) for linear models. Another factor is that this decoder is trained on all subject data, while better performance is achieved with subject specific decoders [5]. Performance might increase with increasing segment length, typically 30-60 s segments are used with linear models.

The convolutional baseline outperforms the decoder baseline by 16.5%. This performance increase might be due to the use of a non-linearity in the final layer and end-to-end training. The dilated model significantly outperforms the convolutional baseline for the same integration window length, possibly due to the bigger capacity and non-linear transformations for both speech envelopes and EEG in the dilated model.

According to Figure 5, the optimal length of the receptive field appears to be between 16 and 32 time samples (250-500 ms). Previous literature shows that this is longer than the optimal length for the linear decoder, which spans from 0 ms to 75-140 ms [5]. A possible explanation is that the more complex and non-linear dilated network can also model some non-linearity of the later auditory EEG responses.

As higher accuracy is achieved in shorter timeframes, the proposed dilated convolutional network has more application potential in a diagnostic setting to measure speech intelligibility, as shorter time is needed for testing. This model can also be incorporated in applications like a smart hearing aid (as part of the feedback loop). Another possible application for the dilated network is auditory attention decoding, in which the attended speaker is identified using only the brain response of the subject.

Some additional strategies might improve this model in future work. Input features with more information of the speech signal might be included (e.g., spectrogram, phonemes, word frequency, etc) to achieve higher performance. As our best model achieves more than 90% accuracy, some ceiling effects may occur. We can compensate for these ceiling effects by shortening the segment length and therefore decreasing overall performance *per segment*.

REFERENCES

[1] Michael J. Crosse, Giovanni M. Di Liberto, Adam Bednar, and Edmund C. Lalor, "The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli," *Frontiers in Human Neuroscience*, vol. 10, 2016.

[2] Edmund C Lalor and John J Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European journal of neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[3] Nai Ding and Jonathan Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11854–11859, July 2012.

[4] Eline Verschueren, Ben Somers, and Tom Francart, "Neural envelope tracking as a measure of speech understanding in cochlear implant users," *Hearing Research*, vol. 373, pp. 23–31, Mar. 2019.

[5] Jonas Vanthornhout, Lien Decruy, Jan Wouters, Jonathan Z. Simon, and Tom Francart, "Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, no. 2, pp. 181–191, Apr. 2018.

[6] Damien Lesenfants, Jonas Vanthornhout, Eline Verschueren, and Tom Francart, "Data-driven spatial filtering for improved measurement of cortical tracking of multiple representations of speech," *Journal of Neural Engineering*, 2019.

[7] Ivan Iotzov and Lucas C. Parra, "EEG can predict speech intelligibility," *Journal of Neural Engineering*, vol. 16, no. 3, pp. 036008, Mar. 2019.

[8] Hassan Akbari, Bahar Khalighinejad, Jose L. Herrero, Ashesh D. Mehta, and Nima Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific Reports*, vol. 9, no. 1, pp. 1–12, Jan. 2019.

[9] Minda Yang, Sameer A Sheth, Catherine A Schevon, Guy M McKhann II, and Nima Mesgarani, "Speech Reconstruction from Human Auditory Cortex with Deep Neural Networks," in *INTERSPEECH*, 2015, p. 5.

[10] Tobias de Taillez, Birger Kollmeier, and Bernd T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, Dec. 2017.

[11] Alain de Cheveigné, Daniel D. E. Wong, Giovanni M. Di Liberto, Jens Hjortkjær, Malcolm Slaney, and Edmund Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, May 2018.

[12] Daniel D. E. Wong, Giovanni M. Di Liberto, and Alain de Cheveigné, "Accurate Modeling of Brain Responses to Speech," *bioRxiv*, p. 509307, July 2019.

[13] Lucas Deckers, Neetha Das, Amir Hossein Ansari, Alexander Bertrand, and Tom Francart, "EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," *bioRxiv*, Dec. 2018, doi:10.1101/475673.

[14] Gregory Ciccarelli, Michael Nolan, Joseph Perricone, Paul T. Calamia, Stephanie Haro, James O'Sullivan, Nima Mesgarani, Thomas F. Quatieri, and Christopher J. Smalt, "Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, Aug. 2019.

[15] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv:1609.03499 [cs]*, Sept. 2016, arXiv: 1609.03499.

[16] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[17] François Chollet et al., "Keras," https://keras.io, 2015.

[18] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, May 2017.

[19] Ben Somers, Tom Francart, and Alexander Bertrand, "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter," *Journal of Neural Engineering*, vol. 15, no. 3, pp. 036007, Feb. 2018.

[20] Tom Francart, Astrid van Wieringen, and Jan Wouters, "APEX 3: a multi-purpose test platform for auditory psychophysical experiments," *Journal of Neuroscience Methods*, vol. 172, no. 2, pp. 283–293, July 2008.